Title: Komputerowe wspomaganie gromadzenia i analizy danych z mikromacierzy oligonukleotydowych w badaniach ekspresji genów

Author: Magdalena Tkacz, Adam Wilczok

Citation style: Tkacz Magdalena, Wilczok Adam. (2008). Komputerowe wspomaganie gromadzenia i analizy danych z mikromacierzy oligonukleotydowych w badaniach ekspresji genów. "Zarządzanie i Technologie Informacyjne" (T. 3 (2008), s. 185-199).

Rozdział 6

KOMPUTEROWE WSPOMAGANIE GROMADZENIA I ANALIZY DANYCH Z MIKROMACIERZY OLIGONUKLEOTYDOWYCH W BADANIACH EKSPRESJI GENÓW

Magdalena Tkacz, Adam Wilczok

Streszczenie: *W pracy przedstawiono tematykę związaną ze stosunkowo nową metodą pozyskiwania informacji w medycynie – technologią mikromacierzy DNA i omówiono przykładowe wykorzystanie możliwości obliczeniowych komputera do filtrowania wyników danych z mikromacierzy. Przedstawiono typy mikromacierzy i etapy eksperymentu wykonywanego z ich wykorzystaniem. Eksperymenty takie umożliwiają m.in. wykrycie zmian w ekspresji genów w żywych komórkach – a co za tym idzie – ich podatność lub odporność na leki i inne ksenobiotyki. Opisano również pojęcia niezbędne do zrozumienia ekspresji genów i ich znaczenie oraz zaproponowano rozwiązanie w postaci hurtowni danych opracowanej z myślą o składowaniu i analizie danych mikromacierzowych.*

Słowa kluczowe: mikromacierz, hurtownia danych, ekspresja genów, analiza danych

Abstract: *A relatively new method of biomedical data acquisition – microarray technology and an example of computational capacity possibilities for filtering of microarray data were described. The types of microarrays used and subsequently microarray experiment steps were discussed. Such kind of experiments makes possible finding changes in gene expression in living cells effected by drugs and other xenobiotics to check the cell resistance or sensibility. Some terms necessary for understanding gene expression and its importance were explained. A proposal of the data warehouse solution, developed with special attention for storage and the analysis of microarray data was presented.*

Keywords: microarray, data warehouse, gene expression, data analysis

1. Wstęp

Druga połowa XX wieku i początek XXI to bardzo szybki rozwój różnego rodzaju technologii, których rozwój pozwala na osiąganie coraz lepszych wyników i badanie różnorakich zjawisk, często wcześniej niedostępnych ze względu na ograniczoną percepcję człowieka. Technologie informatyczne są już również obecne praktycznie we wszystkich dziedzinach życia. Siłą rzeczy, gdy za sprawą rozwoju technologii możliwe jest coraz dokładniejsze opisywanie zjawisk i procesów w otaczającym nas świecie, rośnie też ilość gromadzonych danych. Są to albo dane opisujące coraz dokładniej badane zjawisko lub proces (ze względu na wzrost możliwości pozyskania danych: większą liczbę odczytów w jednostce czasu lub większą dokładność odczytu). Mogą to również być dane pozyskane poprzez wykorzystanie urządzeń zwiększających możliwości percepcji człowieka w odkrywaniu praw rządzących zjawiskami i procesami w naszym otoczeniu, zarówno w skali makro, jak i mikro. Paradoksalnie jednak – większość tak pozyskiwanych danych nie jest użyteczna ze względu na brak możliwości ich wykorzystania: ich ilość przekracza możliwości percepcji i przetworzenia przez człowieka. Ponadto, bardzo często wymagane jest wykorzystywanie wiedzy z różnych dziedzin – właściwie bez wahania można je określić jako interdyscyplinarne. Przedmiotem rozważań mogą być: zarówno wpływ środowiska przyrodniczego, w którym żyjemy, problemy epigenetyki, zrozumienie, w jaki sposób „działa" ludzki organizm, wpływ przyjmowanego pożywienia, wpływ leków oraz ksenobiotyków i inne problemy funkcjonowania żywych organizmów. Czy dysponując coraz bardziej szczegółowymi informacjami na temat funkcjonowania naszych organizmów, będziemy w stanie dokładniej zrozumieć zachodzące w nich mechanizmy? Jedną z metod, która umożliwia i przyspiesza szczegółowe „rozpracowanie" zasad funkcjonowania komórek, a w konsekwencji i całych organizmów jest stosowanie mikromacierzy. Nie jest jednak możliwe wykorzystanie danych pozyskiwanych z mikromacierzy bez wykorzystania technologii informatycznych ze względu na liczbę pozyskiwanych danych i konieczność ich przetworzenia. Systemami przechowywania i przetwarzania danych mogą być zarówno spotykane już powszechnie bazy danych (ang. *databases*), jak i bazy danych z poszerzonymi możliwościami analizy danych określanymi czasami jako hurtownie danych (ang. *data warehouse*). W ostatnich latach dynamicznie rozwija się również dziedzina związana z wydobywaniem informacji ze zgromadzonych danych, tzw. eksploracja/zgłębianie danych (ang. *data exploration*, *data mining*).

Z tymi rozważaniami, związanymi z oceną funkcjonowania żywych komórek (na poziomie molekularnym) z wykorzystaniem technologii mikromacierzy oraz specjalizowanymi systemami przechowywania i przetwarzania danych związany jest ten rozdział.

2. Informacja w organizmach żywych[1]

Procesy zachodzące w żywych komórkach są rezultatem reakcji biochemicznych zachodzących pomiędzy wieloma cząsteczkami, w tym białkami, będącymi swego rodzaju polimerami aminokwasów (czyli stosunkowo prostych związków organicznych). Białka mogą zawierać dodatkowo w swoim składzie określone pierwiastki chemiczne. Liczba aminokwasów w białku może sięgać nawet wielu setek. Białka enzymatyczne biorą udział w produkcji prawie wszystkich związków chemicznych, które można znaleźć w żywych komórkach. Różne białka są produkowane w różnych komórkach, a to, jakie białka są wytwarzane, determinowane jest informacją zawartą w kodzie genetycznym. Każdy z aminokwasów jest kodowany przez specyficzne struktury – nukleotydy, będące częścią dużej cząsteczki, jaką jest kwas deoksyrybonukleinowy (DNA). Przestrzenna struktura DNA to tzw. podwójna helisa zbudowana z czterech monofosforanowych nukleotydów. Całość informacji kodowana jest za pomocą czterech zasad: adeniny, cytozyny, guaniny, tyminy połączonych ze sobą poprzez deoksyrybozy i grupy fosforanowe. Informacja jest kodowana kodonem złożonym z trzech zasad, warunkuje ona lokalizację każdego aminokwasu w białku. Jednostka informacji jest nazywana genem, a liczba genów jest uzależniona od złożoności organizmu – im bardziej złożony organizm, tym w większej liczbie genów zakodowana jest informacja o jego budowie i działaniu. W uproszczeniu, wszystkie geny i DNA zawarte w komórce noszą nazwę genomu. Na gen składają się sekwencje kodujące, tzw. eksony, i sekwencje niekodujące, tzw. introny, co nie oznacza jednak, że introny nie mają wpływu na przekazywanie informacji. Można posłużyć się tu analogią do asynchronicznej transmisji danych w sieci komputerowej – odpowiednikiem sekwencji kodujących są bity danych, odpowiednikami sekwencji niekodujących są bity startu i stopu. Stan aktywności genu określa jego ekspresja. Informacja zawarta w DNA nie może być odczytywana bezpośrednio, musi być najpierw „przepisana" na RNA (kwas rybonukleinowy), z którego, po usunięciu intronów, powstaje mRNA (ang. *messenger* RNA) – tzw. informacyjny RNA. Dopiero z mRNA, w cząsteczce zwanej rybosomem, może zostać odczytana sekwencja kodonów, na podstawie której w procesie zwanym translacją powstanie białko. Oczywiście w procesach tych może dojść do przekłamania informacji. W genach istotnych dla funkcjonowania komórki nie jest ona dopuszczalna i komórka z uszkodzonym genem nie będzie mogła prawidłowo funkcjonować. W takich przypadkach uszkodzenie jest wykrywane, a komórka — niszczona. Wydaje się na podstawie dotychczasowych badań, że w niektórych genach uszkodzenia nie mają większego znaczenia i część informacji może zostać zmodyfikowana. Badania

[1] Passarge 2004; Koolman 2005.

związane z powiązaniem określonych genów z konkretnymi jednostkami chorobowymi są jak najbardziej uzasadnione.

3. Badania genomowe[2]

Jak wspomniano na wstępie, rozwój technologiczny pozwolił na wykonywanie badań w skali, o której niedawno można było tylko pomarzyć. Średnica helisy DNA wynosi ok. 2 nm (od 1,84 dla Z-DNA do 2,55 dla D-DNA), jej długość natomiast może po „rozprostowaniu" przekraczać kilka metrów. Obecnie, dzięki nanotechnologiom możliwe jest prowadzenie badań na pojedynczych niciach DNA. Dzięki mikromacierzom – urządzeniom – czy też może raczej nowej technologii, która pojawiła się w ostatnim dziesięcioleciu XX wieku, możliwy jest odczyt ekspresji genów w poszczególnych tkankach i komórkach (K n u d s e n 2002). Badania takie mogą służyć do określania np. skuteczności działania leku opartego na porównaniu ekspresji genu w komórce przed i po podaniu określonego leku. Jest to wykorzystywane w badaniach nad komórkami nowotworowymi do określenia stopnia skuteczności leków cytotoksycznych i cytostatycznych z równoczesnym uwzględnieniem ewentualnego wystąpienia skutków ubocznych związanych z ich wprowadzeniem do organizmu. Niemal każdy z podawanych związków chemicznych jest w specyficzny sposób metabolizowany w organizmie. Odczyt zmiany ekspresji genu będący odpowiedzią na zastosowany lek może naprowadzić biochemika na trop, który pozwoli na zidentyfikowanie lub odnalezienie konkretnego szlaku przemian metabolicznych, co może wyjaśnić mechanizmy przemian zachodzące w komórce. Badania aktywności genów z użyciem technologii mikromacierzy są bardzo przydatne w ocenie zmian aktywności wielu genów równocześnie (współczesne macierze pozwalają na równoczesną analizę 22 lub 44 tys. pól/genów), ale nie zawsze są precyzyjne. Mogą być punktem startowym do wstępnej selekcji genów, którym należy poświęcić więcej uwagi i przebadać zmianę ich aktywności metodami bardziej precyzyjnymi – hybrydyzacją typu Northern, RT-QPCR i in. Technika mikromacierzy umożliwia ocenę zmiany aktywności genów w czasie, jednak nie umożliwia oceny wzajemnego współoddziaływania genów pomiędzy sobą (gdyż nie wiadomo, czy dodanie substancji A spowoduje równoczesny wzrost aktywności genów x_1 i x_2, czy też substancja A spowodowała wzrost aktywności genu x_1, a ten, produkując enzym, wpłynie na aktywność x_2). Technologia wykonywania mikromacierzy również implikuje powstawanie szumu informacyjnego (migracja sond – oznakowanych oligonukleotydów na sąsiednie pola),

[2] M a z u r c z a k 2004; K o o l m a n 2005.

co dodatkowo utrudnia opracowanie danych uzyskanych podczas eksperymentu. Ze względu na dużą liczbę genów poddawanych równocześnie badaniom nie jest możliwe rozwiązanie powyższych problemów bez korzystania z mocy obliczeniowej komputerów i technologii informatycznych.

4. Mikromacierze DNA i ich zastosowanie

Mikromacierze mogą być wykorzystywane w badaniach dotyczących ekspresji genów, detekcji mutacji oraz genotypowaniu. Generalnie dostępne są dwa typy mikromacierzy: mikromacierze dwukanałowe – zwane też mikromacierzami cDNA i nowsze, jednokanałowe – mikromacierze oligonukleotydowe (np. (affy)). W przypadku mikromacierzy cDNA równoczesnej analizie podlegają zarówno kontrolne (zdrowe) próbki tkanek, jak i badane, w macierzach oligonukleotydowych – każdorazowo jeden typ tkanki (wiki). Mikromacierze cDNA są tańsze w wykonaniu – znakowane fluorochromami próbki nici cDNA nanosi się na szklaną lub plastikowa płytkę. Ich wykonanie może być realizowane bezpośrednio w ośrodku wykonującym badania – wyposażonym w odpowiednie zautomatyzowane stanowisko zapewniające zachowanie dużej precyzji w umieszczaniu tysięcy genów na powierzchni jednego cala kwadratowego. Mikromacierze GeneChip są droższe, do wykonania badania konieczny jest zakup konkretnego mikrochipa, czyli płytki z odpowiednio przygotowanymi oligonukleotydami umieszczonymi na płytce. Dokładniejsze informacje dotyczące mikromacierzy można znaleźć w (K i s i e l i in. 2004) i (F r i e n d, S t o u g h t o n 2002). W (biodav) można obejrzeć animacje Flash przedstawiającą zasadę wykonywania eksperymentu z mikromacierzami cDNA. W 2004 roku Carole L. Yauk i współpracownicy (Yauk) porównali 6 różnych technologii wytwarzania i wykorzystania mikromacierzy. Opierając się na standardzie MIAME (ang. Minimal Information About a Microarray Experiment) (miame), podkreślili, że niewiele jeszcze wiadomo o możliwościach interpretacji profilu transkrypcji i różnicowej ekspresji badanych genów. Porównując ekspresję genów na platformach oligonukleotydowych i platformach cDNA, autorzy ci porównali powtarzalność uzyskanych wyników, wykazując, jak wiele zależy od jakości wykonania samej platformy.

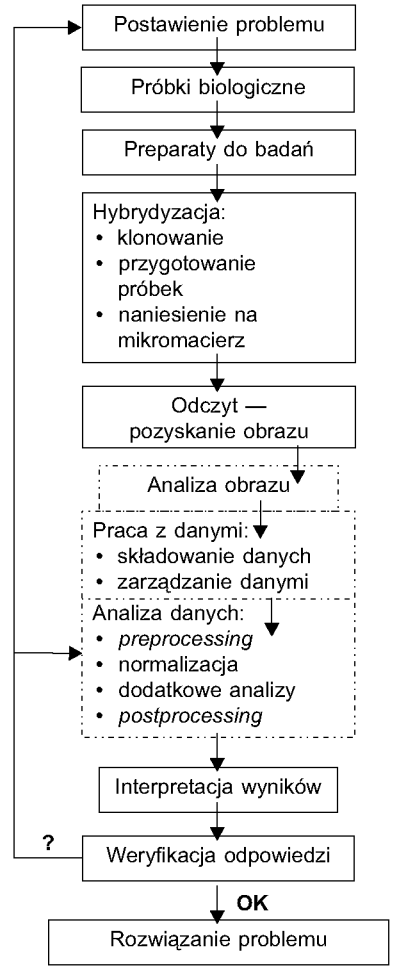
Jednym z często przeprowadzanych badań z wykorzystaniem tej technologii są badania nad skutecznością leków, np. przeciwnowotworowych. Badanie polega na pomiarze ekspresji genów w tkankach prawidłowych, tkance nowotworowej – i potem na pomiarze ekspresji genów po dodaniu różnych substancji leczniczych. Ekspresja genów może być zmniejszona lub zwiększona – w zależności od działania leku. Z drugiej strony mikromacierze mogą wskazać na

obecność lub aktywność genów, które są ściśle związane z mechanizmami działania badanych substancji, np. o działaniu przeciwnowotworowym. Należy brać pod uwagę, że dostępne w handlu matryce mikromacierzy z konieczności mają „z góry" ustaloną ilość DNA (czy oligonukleotydu) na nośniku (nylonowym lub na płytce szklanej). Przy zbyt małym ich stężeniu można oczekiwać nawet ekstremalnie niskich poziomów ekspresji, pomimo iż „teoretycznie" dany gen jest szczególnie wrażliwy na działanie określonego związku chemicznego.

Opisane powyżej zjawisko było podstawą rozwoju badań nad ulepszeniem technik mikromacierzy. W wielu przypadkach opracowano metody, gdzie na powierzchni siatki mikromacierzy umieszczono kilkanaście lub kilkaset genów, których właściwości były albo poprzednio opisane, albo o których wiedziano, że mogłyby uczestniczyć w mechanizmach działania badanych substancji, gdyż znane czynniki przeciwnowotworowe, np. o charakterze antymetabolitów, są stosowane w praktyce klinicznej od wielu lat. Dużą wagę przypisuje się ustaleniom wielkości stosowanej dawki u pacjentów z określoną chorobą nowotworową, pojawianiem się efektów ubocznych, wyborem mieszaniny leków cytostatycznych najskuteczniejszych w określonych przypadkach nowotworów, ustaleniem sposobu i częstości podawania preparatów itd. Efekty uboczne wielu leków stosowanych w chemioterapii nowotworów urosły do miary poważnego problemu naukowego i etycznego w świetle artykułu opublikowanego w „Nature Medicine" z 23 lipca 2006 roku (K e r k e l i in. 2006) dotyczącego kardiotoksyczności w przypadkach stosowania leku Gleevec w terapii nowotworowej. Z badań prowadzonych w wielu ośrodkach akademickich wynika, że znany jako lek inhibitor kinazy tyrozynowej imatinib (Gleevec, Novartis), stosowany w leczeniu białaczki szpikowej wywołał u 10 osobników poważne zaburzenia czynności serca. Ostatnie doniesienia, dotyczące pojawienia się ciężkich schorzeń serca u ludzi przyjmujących Gleevec, zobowiązują do dokładniejszej analizy mechanizmów zwalczania chorób nowotworowych opartych na efektach związanych z działaniem kinaz. Jest nie do zaakceptowania fakt, że pacjent co prawda zostanie „wyleczony" z nowotworu, ale umrze z powodu choroby serca wywołanej przyjmowaniem leków przeciwnowotworowych.

5. Przeprowadzanie eksperymentów z wykorzystaniem mikromacierzy

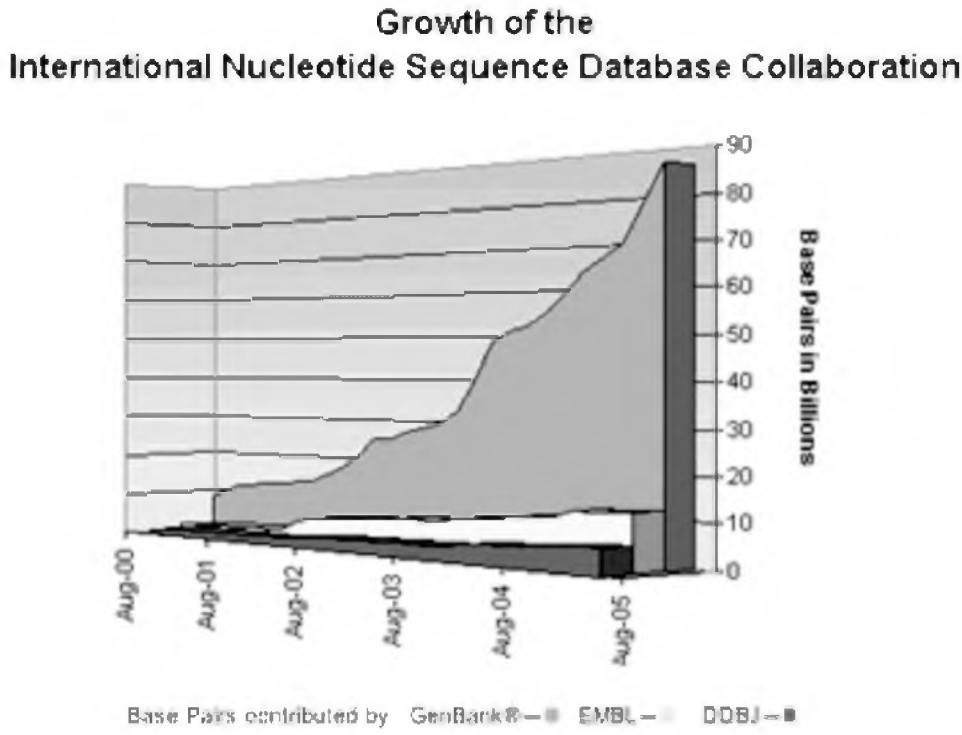
Ogólny schemat przebiegu eksperymentu z wykorzystaniem technologii mikromacierzy przedstawiono na rys. 1. Linią przerywaną oznaczono etapy, w których pomoc i wsparcie informatyków są niezbędne.



Rys. 1. Przebieg eksperymentu z wykorzystaniem technologii mikromacierzy DNA

Początkowy brak ustalonego standardu odnośnie do przeprowadzania eksperymentów z wykorzystaniem mikromacierzy spowodował pewne trudności związane ze współpracą oraz wymianą informacji. Obecnie został zaakceptowany przez wiele ośrodków standard oparty na XML, wspomniany wcześniej MIAME, który określa minimum informacji, jakie powinien zawierać opis badania-doświadczenia z wykorzystaniem mikromacierzy, by umożliwić wykorzystanie wyników szerszej społeczności badaczy.

Jedną z większych baz danych, zawierających informacje dotyczące poszczególnych sekwencji, dostępną dla wszystkich zainteresowanych, jest GenBank (genbank), a wzrost liczby danych w bazie przedstawiono na rys. 2. Dodatkowe, znane bazy dostępne *online* to np. EMBL (embl) oraz DDBJ (ddbj).



Rys. 2. Wzrost rozmiaru bazy danych GenBank (za: genbank).

Nietrudno zauważyć, że wzrost liczby danych – szczególnie w przypadku GenBank – ma już tendencję wykładniczą. Jest rzeczą oczywistą, że przy tak ogromnej, już zgromadzonej, ilości danych – i stałym jej wzroście w takim tempie nie jest możliwa analiza i skorzystanie z tych danych bez wyspecjalizowanych systemów informatycznych. Systemy te powinny być zaprojektowane ze szczególną starannością, jeśli chodzi o dobór algorytmów oraz optymalizacje kodu pod względem wykorzystania pamięci operacyjnej, co powinno umożliwić optymalną pracę systemów wspomagających analizę danych tego typu. Nie bez znaczenia podczas projektowania systemów dedykowanych do pracy w tym zakresie będzie uwzględnienie intensywnej wymiany danych z nośników danych. Często przydatne jest też pobieranie dodatkowych informacji poprzez sieć Internet. Powyższe uwagi z góry wymuszają przyjęcie określonych założeń projektowych odnośnie do bioinformatycznych systemów komputerowych – mających umożliwiać gromadzenie i analizę danych biologicznych lub medycznych.

6. Przygotowanie próbek i pozyskiwanie danych

W badaniach techniką mikromacierzy można wyróżnić kilka etapów. Część związana z odpowiednim przygotowaniem próbek z tkanek lub komórek jest stosunkowo czaso-, praco- i kosztochłonna. Ta część eksperymentu wykonywana jest przez osoby o przygotowaniu medycznym, biologicznym, farmaceutycznym lub innym, ale najczęściej niezwiązanym bezpośrednio z informatyką. Dostępna aparatura umożliwia odczyt danych, np. poziomu fluorescencji z poszczególnych mikromacierzy. Paradoksalnie jednak, postęp naukowy i technologiczny umożliwiający pozyskiwanie coraz dokładniejszych i coraz większych ilości danych właściwie tylko pozornie zwiększył możliwość ich wykorzystania w praktyce, gdyż liczba sukcesywnie pozyskiwanych i składowanych danych jest ogromna. W chwili obecnej w większości nauk przyrodniczych zauważalny staje się – znany już od kilku lat środowisku informatycznemu – dylemat związany z ilością informacji i koniecznością jej przetworzenia i/lub wyszukania. Najbardziej dobitnym przykładem są zasoby sieci WWW: problemem jest nie brak informacji, ale dotarcie do poszukiwanej, wiarygodnej informacji i znalezienie właściwych odnośników do konkretnej tematyki w rozsądnym czasie. Spowodowało to (zauważalny od kilku lat) wzrost podejmowanych wysiłków (prac naukowych i tematów badawczych) zmierzających do opracowania i wdrożenia różnego rodzaju metod selekcji i wyszukiwania informacji.

W przypadku danych z mikromacierzy – pomijając zagadnienia związane z błędami pomiarowymi, wykonaniem badania (włącznie z dostępem do materiału do badań) – problemem staje się wyszukanie interesującej nas w danym momencie informacji pozwalającej na uzyskanie odpowiedzi dotyczącej określonego problemu biologicznego lub biomedycznego. Informatycy nie są w stanie wykonywać omawianych badań i/lub interpretować uzyskanych wyników. Mogą jednak pomóc w opracowaniu skutecznych algorytmów i metod, czy też zaprojektowaniu i wdrożeniu systemów informatycznych (bazodanowych, hurtowni danych, systemów z elementami heurystyki i elementów inteligencji obliczeniowej) ułatwiających przetworzenie i wydobycie użytecznych z praktycznego punktu widzenia informacji. Nie można wykluczyć, że konieczne będzie opracowanie nowych algorytmów i metod, ale być może wystarczy efektywne wykorzystanie lub adaptacja znanych już i stosowanych sposobów oceny aktywności badanych genów.

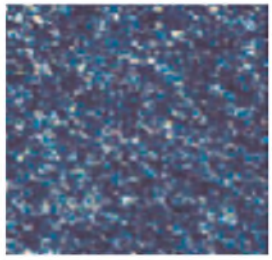
7. Obróbka i analiza danych z mikromacierzy

W przeprowadzonych przez autorów eksperymentach i przykładowej analizie danych jako ich źródło wykorzystano mikromacierz oligonukleotydową Affymetrix. Fragment uzyskanych danych jest prezentowany w postaci tabeli (zob. tab. 1), ale może być on również uzyskany w postaci obrazu rastrowego, w którym poziom ekspresji określonego genu odpowiada kolorowi na obrazie (rys. 3).

Tabela 1

Przykładowe dane z mikromacierzy w postaci liczb (ekspresja genów)
(Liczby oznaczają intensywność fluorescencji dla poszczególnych genów (kolumna 1)
w trzech hodowlach komórek (kolumny 2 do 4))

Symbol genu	Hodowla kontrolna komórek	Hodowla kontrolna + związek A	Hodowla kontrolna + związek B
MMP14	5.904916	5.859608	5.949325
SPARC	2.454956	2.571326	2.367730
EPAS1	3.788308	3.904229	3.756821
EPAS1	3.481430	3.238905	3.437383
BTG1	5.457769	5.828565	7.112901



Rys. 3. Przykładowy obraz z mikromacierzy – jedna hodowla

W obu przypadkach, ze względu na dużą liczbę danych (np. ok. 22 tys.) dla każdego odczytu, zarówno w postaci tabeli z danymi, jak i obrazu, konieczne jest wykorzystanie mocy obliczeniowej komputerów do wykonania tego zadania. Przykładowo, dla danych w tab. 1 należy wyszukać różnice, ewentualnie powtarzalne wzorce w wielu próbkach i/lub zależności dla 66 tys. odczytów, czyli po dodatkowym uwzględnieniu kolumny z opisem daje to 88 tys. danych jednostkowych do przetworzenia. Na szczęście, wraz z rozwojem technologii badawczych w naukach przyrodniczych rosną także możliwości obliczeniowe komputerów, gdyż zgodnie z zasadą Moore'a moc obliczeniowa komputerów podwaja się co 18 miesięcy. Co prawda, coraz częściej słyszy się opinie, że technologia wykonywania procesorów spowoduje, że zasada Moore'a przestanie obowiązywać,

ale pojawiają się w tym kontekście takie sformułowania, jak komputery kwantowe czy też nanotechnologie – więc być może zasada Moore'a będzie jeszcze obowiązywać przez kilka lat. Oczywiście nie zwalnia to informatyków od opracowywania skutecznych, szybkich i wydajnych algorytmów do obliczeń: jeśli moc obliczeniowa komputera wzrośnie dwukrotnie, to wydajny algorytm umożliwi przetworzenie większej liczby danych w tym samym czasie. Złożony algorytm nie spowoduje jednak znaczącej poprawy wydajności. W zakresie badań genetycznych, związanych z doborem leków, szybkość działania algorytmu może być parametrem krytycznym (przy założeniu, że dobór leku jest faktycznie wykonywany na podstawie badania ekspresji genów). Jest prawie pewne, że nie bez znaczenia będzie fakt, czy lek zostanie dobrany powiedzmy po tygodniu, czy po kilku godzinach obliczeń.

Do analizy wyodrębnienia z mikromacierzy genów o różniącej się ekspresji wykorzystywane są różne metody. Zdecydowanie przydatne byłoby tu opracowanie powszechnie zaakceptowanej, jednolitej metodyki postępowania. Porównywanie wyników uzyskanych różnymi metodami obróbki danych, gdy zbiory uzyskane po klasyfikowaniu/grupowaniu genów przy użyciu nawet tych samych algorytmów grupowania, ale różnych miar i metryk, jest praktycznie niemożliwe. Na dzień dzisiejszy do analizy genów są stosowane metody statystyczne (co jest zasadne przy odpowiednio dużej liczbie badanych prób), metody oparte na redukcji wymiarów przestrzeni rozważań (np. PCA – ang. Principal Component Analysis – metoda składowych głównych), by uniknąć „przekleństwa wymiarowości", metody klastrowania danych, metody sztucznej inteligencji – sieci neuronowe samoorganizujące się i inne metody znane w eksploracji danych. Przy relatywnie niewielkiej liczbie mikromacierzy do analizy mogą okazać się przydatne powszechnie dostępne aplikacje, które umożliwiają zdefiniowanie prostych kryteriów, według których można wykonywać filtrację danych, np. arkusze kalkulacyjne, takie jak komercyjny Excel z pakietu Microsoft Office, Calc

Tabela 2

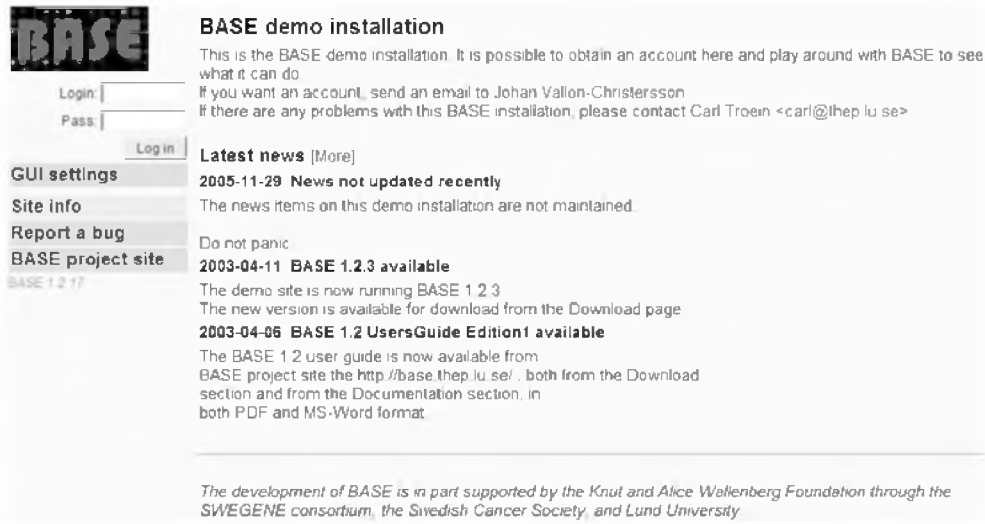
Przykładowe, wyfiltrowane z odczytów z mikromacierzy geny o zmniejszonej ekspresji po podaniu związku A

				Ile-genów ->	4		
		max_różnicy=	1.5328	**45.5**	**25**	769	947
				max^	próg %	ile^	ile^
Symbol genu	Hodowla kontrolna + związek A	hodowla kontrolna + związek B	różnica	% kontr	25	poniżej	powyżej
PLOD2	3.2776	4.6250	–1.3474	41.1	1	**1**	0
EREG	3.9348	5.7245	–1.7897	45.5	1	**1**	0
KLRC1	2.9384	3.9580	–1.0196	34.7	1	**1**	0
APPL	3.3964	4.6719	–1.2755	37.6	1	**1**	0

z Open Office czy – również darmowy – Gnumeric, który dodatkowo umożliwia przeprowadzenie prostych analiz statystycznych i tworzenie modyfikowalnych wykresów w większym zakresie niż w Excel. Przykładowe, przefiltrowane w arkuszu kalkulacyjnym dane pokazano w tabeli 2.

8. Technologie informatyczne a mikromacierze DNA

Obecnie rozwijany jest (na licencji GNU GPL) BASE (ang. BioArray Software Environment) (base) – projekt systemu bazodanowego – hurtowni danych umożliwiający gromadzenie danych pozyskanych z mikromacierzy, zarządzanie nimi oraz ich analizę. System ten wykorzystuje technologie internetowe, gdzie interfejsem użytkownika jest przeglądarka stron WWW, a wszystkie operacje wykonywane są na serwerze. W obecnie dostępnej stabilnej wersji systemu możliwe jest gromadzenie i analiza danych z mikromacierzy cDNA. Wersja BASE2 (rys. 4), dostępna na razie w fazie testowej, obsługuje również dane otrzymywane z mikromacierzy oligonukleotydowych – GeneChip Affymetrix. System BASE jest stale rozwijany i wyposażany w dodatkowe moduły zwiększające jego możliwości. Ze względu na licencję (GNU GPL) może być użyty w innych jednostkach naukowych jako platforma dla analizy ekspresji genów z wykorzystaniem mikromacierzy.



Rys. 4. Serwer demonstracyjny BASE (base-demo)

Po zalogowaniu do systemu zmianie ulegają opcje menu (rys. 5), umożliwiając użytkownikowi przetestowanie możliwości oprogramowania na przykładowych, dostępnych w systemie danych.

Rys. 5. System BASE po zalogowaniu (base-demo)

9. Podsumowanie

Współczesne nauki przyrodnicze dysponują ogromnym arsenałem urządzeń umożliwiającym prowadzenie eksperymentów i gromadzenie danych. Brak jest jednak opracowanej spójnej metodologii pozwalającej na analizę i obróbkę tak dużych zbiorów informacji, umożliwiającej dokonywanie obiektywnych i „bezpiecznych" porównań wyników. Być może zadanie to zostanie wykonane przez bioinformatyków w ramach rozwoju bioinformatyki (której zadaniem może być m.in. pomoc w analizie i obróbce danych z mikromacierzy), a która jest rozwijającą się obecnie dyscypliną nauk obliczeniowych – określonych w (mnisw) jako „pomost pomiędzy dyscyplinami teoretycznymi i eksperymentalnymi". Nauki obliczeniowe z kolei są wymienione jako te, które poprzez „Symulacje i inne obliczenia komputerowe pozwalają unikać zazwyczaj bardziej kosztownych i czasochłonnych eksperymentów rzeczywistych". Opracowanie takich metod pozwoliłoby na szybkie, dokładne i prawie pozbawione błędów dobieranie leków w przypadku różnych chorób. Uzyskiwane wyniki mogłyby być wykorzystywane w medycynie i biologii molekularnej.

Wykorzystywanymi technologiami informatycznymi będą hurtownie danych, klastry lub farmy serwerów wyposażone w starannie opracowane i wyselekcjonowane pod odpowiednim kątem algorytmy i rozwiązania. Niewątpliwie opracowanie efektywnego systemu przetwarzania tak dużej liczby danych w celach diagnostycznych i terapeutycznych w medycynie jest w tej chwili jednym z ważniejszych wyzwań. Może to stanowić przedmiot rozważań i prac zarówno algorytmików, kryptografów i kryptologów, specjalistów zajmujących się metodami inteligencji obliczeniowej, ekspertów od baz i hurtowni danych, osób zajmujących się analizą obrazów, administratorów systemów informatycznych i osób zajmujących się sieciami komputerowymi i optymalizacją przepływów informacji w sieciach.

Dodatkową trudnością jest fakt, że aby opracować efektywne systemy, niezbędna jest współpraca specjalistów różnych dziedzin pracujących zespołowo. Informatyk nie będzie w stanie stworzyć adekwatnego modelu bez pomocy przyrodnika (biologa, biochemika, farmaceuty), a po stworzeniu modelu zinterpretować i zweryfikować otrzymanych wyników, natomiast przyrodnik nie będzie w stanie uzyskać zadowalających wyników ze względu na liczbę danych, które należy przetworzyć do ich uzyskania. Złożoność problemów oraz liczba danych, które należy przetwarzać, będą wymagać zapewne dostosowania już istniejących lub wręcz opracowania i wdrożenia nowych, efektywnych algorytmów i rozwiązań umożliwiających przetwarzanie i wykorzystywanie takiej liczby danych.

Literatura

Friend S., Stoughton R., 2002: *Magiczne mikromacierze*. „Świat Nauki", 4.

Kerkel R., Grazette L., Yacobi R., Iliescu C., Patten R., Beahm C., Walters B., Shevtsov S., Pesant S., Clubb F.J., Rosenzweig A., Salomon R.N., Van Etten R.A., Alroy J., Durand J.B., Force T., 2006: *Gleevec cardiotoxicity of the cancer therapeutic agent imatinib mexylate*. „Nature Medicine", 12: 908–916.

Kisiel A., Skąpska A., Markiewicz W.T., Figlerowicz M., 2004: *Mikromacierze DNA*. „Kosmos" 53/3–4: 295–303. kosmos.icm.edu.pl/PDF/2004/295.pdf.

Knudsen S., 2002: *A Biologist's guide to analysis of DNA microarray data*. New York: Wiley & Sons.

Koolman J., Rohm K.H., 2005: *Biochemia. Ilustrowany przewodnik*. Węglarz L., Wilczok T. (red., tłum.). Warszawa: PZWL.

Ministerstwo Nauki i Szkolnictwa Wyższego. *Priorytetowe kierunki badawcze*. http://www.nauka.gov.pl/mein/index.jsp?news_cat_id=79&news_id=363&layout=2&page=text&place=Lead01.

Passarge E., 2004: *Genetyka. Ilustrowany przewodnik*. Mazurczak T. (red., tłum.). Warszawa: PZWL.

Yauk C.L., Berndt M.L., Williams A., Douglas G.: *Comprehensive comparison of six microarray technologies*. „Nucleic Acid Research" 32(15): 124. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=516080.
base.thep.lu.se/.
en.wikipedia.org/wiki/DNA_microarray.
http://base1.thep.lu.se/demo.
http://www.mged.org/Workgroups/MIAME/miame.html.
http://www.ncbi.nlm.nih.gov/Genbank/index.html.
www.affymetrix.com.
www.bio.davidson.edu/Courses/genomics/chip/chip.html.
www.ddbj.nig.ac.jp/.
www.embl-heidelberg.de/.
www.wiw.pl/biologia/Genetyka/JezykGenow/Esej.asp?base=r&cp=1&ce=23.